Penultimate draft. Final version can be found at https://link.springer.com/article/10.1007/s11016-020-00539-7

Big Data in the Experimental Life Sciences

Bruno J. Strasser, *Collecting Experiments: Making Big Data Biology*, Chicago: The University of Chicago Press, 392 pp., $ 45.00

Emanuele Ratti

mnl.ratti@gmail.com

Reilly Center for Science, Technology, and Values

University of Notre Dame

Notre Dame, IN, USA

Bruno Strasser's *Collecting Experiments* is an essential book to understand the present 'big data' situation in the experimental life sciences. It provides a solid framework to interpret the relationship between collecting data and experimenting in biology, it thoroughly shows how we got to the present 'big data' biology situation, and it articulates suggestions on how we should interpret bioinformatics as a discipline and its relation to experimental biology, even though this is not his central goal. But it should not be read in isolation. Rather, it has to be considered in tandem with other important scholarly works, such as, but not limited to, November's *Biomedical Computing* (2012), Leonelli's *Data-Centric Biology* (2016), and Stevens' *Life Out of Sequence* (2013).

What I intend to do in this review is to provide (1) some context to understand the book, (2) a summary of the main contents with emphasis on the important contributions that it provides, and (3) two suggestions for future work.

The Context

Strasser's book is motivated by an urgent need to understand the big data revolution in biology. The significance of the so-called 'data-driven', or 'data-intensive' turn in biology has been widely debated, in particular its epistemological magnitude, and the changes in how biologists perceive their profession. But, according to Strasser, in order to understand what happened in the past 30 years or so we should take a step back, and realize how the interplay between databases and molecular biology is only an instance of a general dynamics affecting the experimental life sciences in the 20th century. Therefore, Strasser's book "is about the development and use of data collections in the experimental life sciences from the early twentieth century to the present" (6).

A work on the role of collections in the experimental life sciences is important and timely because something about collecting and experimenting has been perceived throughout the history of biology as in sharp opposition, as if there was a science war between the two. Strasser lists some ways in which this opposition has been conceptualized. For instance, the opposition has been understood as a tension between natural history disciplines such as taxonomy, paleontology, etc., and experimentalists since the mid-19th century. Others tried to make sense of the opposition by building it around the 'laboratory' against the 'museum'. But Strasser seems to identify the real opposition in something even more fundamental, which is the value of biological diversity and comparisons versus the narrow focus or 'exemplarism' of many of the

experimentalists such as those working on model organisms. He characterizes this opposition as a tension between two 'ways of knowing,' "the comparative and the experimental, the former centered on collections and the latter on exemplary systems" (16). Strasser refers explicitly to Pickstone's ways of knowing, and he stresses that these are analytical rather than taxonomical categories: comparing and experimenting are "the ingredients, coexisting in different proportions" in the life sciences of the 20th century (16). Against the 'opposition narrative', Strasser's book is about how these two ways of knowing somehow interacted and most of the times hybridized in the life sciences in the 20th century. Therefore, this book "is not about the clash of scientific disciplines or research fields (natural history against molecular biology), but about the historical dynamics of their epistemic components (comparing and experimenting)" (16). The same dynamical interactions between experimenting and comparing is at play in the so-called 'big data' biology.

The Book

Strasser starts the story of the dynamics between comparing and experimenting by reconstructing the importance of collections of organisms for those community of experimentalists that were genuinely 'exemplarists', such as those who worked on selected model organisms. The important point that Strasser wants to make in Chapter 1 is that stock collections have been for experimentalists in the life sciences what museums have been for naturalists. In particular, they played the role of repositories of organisms, centers of standardization, centers of distribution, tools for research, centers for coordination, and even as institutions defining social and epistemic norms. Ingredients of the 'comparing' way of knowing materialized in these stock collections by

providing “a standardized and stabilized nature for scientists to study” (64), even though classification *per se* was not a concern, unlike in natural history.

Chapter 2 is about those who promoted the ‘experimentalization’ of taxonomy. Strasser reconstructs how species came to be studied at the biochemical level, in particular by finding ways to analyze blood and its many components. What is striking about this chapter is that the transformation of taxonomy via biochemical analysis was motivated as a way to amend and rectify the subjectivism intrinsic to taxonomical studies based on morphology. In other words, studying species at the molecular level provided a mean to ‘quantify’ and make more objective the way biologists classified species. ‘Objectivity’ was a virtue held in high regard by experimentalists. This is a case of epistemic virtues of experimentalists being transferred to the context of natural history. Even though “epistemic practices remained largely those of naturalists: they were still collecting and comparing, although in the form of experimental data instead of bones and fossils” (107).

In Chapter 3, Strasser analyzes the other side of the coin: how experimentalists hybridized with a comparative culture, in particular in biochemistry. He describes in detail the achievements of Margaret Dayhoff who created the first sequence database. The *Atlas of Protein Sequence and Structure*, which was published at first in book format in 1965, paved the way to electronic databases. Dayhoff positioned her Atlas firmly within the experimentalist tradition. Strasser lists some of the discoveries that Dayhoff was able to make using the Atlas, thereby showing the importance of comparative approaches for experimentalists in generating hypotheses that could be explored in the laboratory. According to Strasser, the Atlas exemplifies a hybrid practice: “the production of knowledge through the collection, comparison, and computation of data rested on the comparative and experimental ways of knowing” (153).

In Chapter 4, Strasser delineates the history of an analogous collection for crystallographers, this time of protein structures: the famous *Protein Data Bank* (PDB), published in 1971. PDB reflected the importance and the impact of three-dimensional computer graphics and computer networks in experimental crystallography. These technological tools offered clear advantages with respect to the typical 'physical models' of protein structures: manipulating virtual models rather than building actual physical models made modeling a much more effective practice. Moreover, virtual models enhanced the process of identifying the position of each atom in the structure – "with physical models, researchers had to tediously measure the relative position of each atom with a ruler" and "interactive possibilities … allowed researchers to 'see' and 'manipulate' molecules as if they were real objects" (160). Therefore, PDB is one example of a collection that became a tool to generate experimental knowledge, and not just a repository of data. But PDB was also used as a taxonomic tool, "for bringing order into the great diversity of structures it contained … hierarchical categories, families, and superfamilies, following a standard taxonomic practice" (176). In other words, PDB represents a hybridization of the two ways of knowing. In fact, PDB "developed as an institution for the preservation and distribution of knowledge … [and] it became an instrument for the production of new knowledge" alongside laboratory instruments (191).

Chapter 5 reconstructs the history of the development of GenBank, which paved the way to a new set of theoretical research practices based on comparison and computation that came to be known as *bioinformatics* or *computational biology*. GenBank is an interesting case study for a number of reasons. First, it is a paradigmatic example of how databases are not mere repositories, but rather tools for producing knowledge. Moreover, GenBank is interesting because it was created as a public, open, and free resource even if it was developed in the context

of molecular biology in which there was, at the time, a push to patent almost everything. Even if researchers have been reluctant to share data at first, as soon the right system of incentives was provided data started to flow. GenBank reflects two major transformations in the experimental life sciences at the end of the 20th century. On the one hand, the changing moral economies of the rise of open access, and the "changing research practices made possible by electronic databases (the rise of comparative practices)" (224). In Chapter 6, issues of open science and how this impacted the ethos of experimentalists are discussed in much more detail.

Suggestions

I want to conclude this review by highlighting two aspects of the big data biology landscape that the book does not describe in detail. Nonetheless, the book offers a valid starting point to develop them. This is not a criticism of the book. Rather, given the complexity of the transformations in the experimental life sciences in the 20th century, it is impossible to cover all possible angles.

First, there is the relation between computer science and biology. Strasser describes in great detail how computational infrastructures and computational tools have facilitated practices where comparing and experimenting hybridized creating a new culture, and even a new discipline, bioinformatics. But sometimes I had the impression that computer science comes out of nowhere. In other words, computer science seems to be a mere tool, out there, for biologists to use. For this reason, I had the impression that what Strasser is describing is not a real synthesis between computer science and biology, but rather a mere instrumental use of the former by the latter. But to use Stevens' words, "the computer [in biology] brought with it epistemic and

institutional reorganizations" (Stevens 2013, 39). I look forward to reading and maybe researching more about this. Moreover, it should also be noted that an impression of 'instrumental relation' rather than real synthesis, sometimes emerges also in the way Strasser describes the dynamics between comparing and experimenting. Strasser describes the relation between the two ways of knowing "as the emergence of a 'hybrid culture' rather than as the domination of one culture over the other" (107). However, at times the nature of this hybridization was not entirely clear to me. How are we supposed to understand this integration? Are the cases described by Strasser examples of interdisciplinary, multidisciplinary, or transdisciplinary practices? Is there a real synthesis, or it is just practitioners from one discipline using instruments or importing epistemic values from another?

The second aspect about which I would like to hear more is how to characterize experimentalism. This is not a petty question, because it has consequences for how we think about bioinformatics. If the comparative and the experimental are analytic rather than taxonomical categories, then we should find different combinations of these two in the different scientific cultures that Strasser describes. In the molecular biology of the end of the 20th century the experimental ingredient is the set of practices typical of the experimental life sciences used to materially manipulate biological entities, while the comparative part is the use of databases. But what about *bioinformatics*? It seems to me that Strasser emphasizes the comparative side of bioinformatics more than its experimentalist side. He says that the collection "of standardized biological data in electronic format opened the door for the development of sophisticated algorithms to analyze it … research *in silico* was becoming a legitimate way of producing knowledge about nature" (253). If bioinformatics is a hybrid between the ways of knowing of the comparative and the experimental, and the comparative ingredient is clear, where is exactly the

experimental? Is it just the fact that it aids experiments, or that it can perform experiments of some sort?

Determining the experimental ingredient of bioinformatics is complicated, because bioinformatics is a term that covers a broad range of practices and professional figures. As Stevens noted (2013), there are at least two types of individuals working in bioinformatics. First, there are those who have been trained as wet-lab biologists, and they have learned how to code later. These individuals aim at solving biological problems by training and using algorithms in addition to wet-lab procedures. But there are also individuals who have been trained in computer science, and they are interested in solving computer science problems in the biological domain, such as how to speed up a computational process, how to write portable codes, etc. In the latter case, computer scientists are instrumental to biologists. In the former case, this is where bioinformatics as a Strasser-type hybrid between comparing and experimenting emerges. But what kind of experiments do bioinformaticians do? The answer to this question depends on what experiments are. In Strasser's book, as an analytic, and not a taxonomic, category, 'experimental' covers a broad range of practices: not just "manipulations intended to uncover *causal* mechanism" (14), but also "results as different as microscopic observations of cells and DNA sequence data, all produced through the manipulation of nature, usually in the laboratory, with specialized instruments" (14).

'Manipulation' seems to be the keyword here. Together with Federico Boem, I have articulated the idea that data mining practices are forms of manipulation of data, even though they are not 'material' manipulation (Boem and Ratti 2016). We suggest that practitioners in bioinformatics 'manipulate' data sets with computational tools, via *abstraction* and *idealization*. In analogous ways molecular biologists manipulate biological entities, even though not in the

same ‘material’ way. While I now think that our proposal was weak, it still constitutes a way in which we can think about the experimental ingredient in bioinformatics. But this is not the only way. For instance, consider the debate about computer simulations and experiments, and to what extent and from which point of view these practices are similar. A viable way to delineate the ‘experimental’ ingredient in bioinformatics may be to consider the different angles through which computer simulations and experiments have been compared, and see if some shed light on the relation between machine learning and other techniques in bioinformatics and experimentation.

References

Boem, Federico, and Emanuele Ratti. 2016. “Towards a Notion of Intervention in Big-Data Biology and Molecular Medicine.” In *Philosophy of Molecular Medicine: Foundational Issues in Research and Practice*, edited by Giovanni Boniolo and Marco Nathan, 147–64. London: Routledge.

Leonelli, Sabina. 2016. *Data-Centric Biology*. Chicago: University of Chicago Press.

November, Joseph. 2012. *Biomedical Computing - Digitizing Life in the United States*. The Johns Hopkins University Press.

Stevens, Hallam. 2013. *Life out of Sequence - A Data-Driven History of Bioinformatics*. Chicago: Chicago University Press.